Relations Between Random Coding Exponents and the Statistical Physics of Random Codes *

Neri Merhav

February 1, 2008

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, ISRAEL

Abstract

The partition function pertaining to finite–temperature decoding of a (typical) randomly chosen code is known to have three types of behavior, corresponding to three phases in the plane of rate vs. temperature: the *ferromagnetic phase*, corresponding to correct decoding, the *paramagnetic phase*, of complete disorder, which is dominated by exponentially many incorrect codewords, and the *glassy phase* (or the condensed phase), where the system is frozen at minimum energy and dominated by subexponentially many incorrect codewords. We show that the statistical physics associated with the two latter phases are intimately related to random coding exponents. In particular, the exponent associated with the probability of correct decoding at rates above capacity is directly related to the free energy in the glassy phase, and the exponent associated with probability of error (the error exponent) at rates below capacity, is strongly related to the free energy in the paramagnetic phase. In fact, we derive alternative expressions of these exponents in terms of the corresponding free energies, and make an attempt to obtain some insights from these expressions. Finally, as a side result, we also compare the phase diagram associated with a simple finite–temperature universal decoder, for discrete memoryless channels, to that of the finite–temperature decoder that is aware of the channel statistics.

Index Terms: random coding, free energy, partition function, random energy model (REM), phase transitions, error exponents.

*This research is partially supported by the Israel Science Foundation (ISF), grant no. 223/05. Part of this work was carried out during a visit at HP Laboratories, Palo Alto, CA, U.S.A. in the Summer of 2007.

1 Introduction

In the last few decades it has become apparent that many problems in Information Theory, and the channel coding problem in particular, can be mapped onto (and interpreted as) analogous problems in the area of statistical physics of disordered systems (such as spin glass models). Such analogies are useful because physical insights, as well as statistical mechanical tools and analysis techniques (like the replica method), can be harnessed in order to advance the knowledge and the understanding with regard to the information–theoretic problem under discussion. A very small, and by no means exhaustive, sample of works along this line includes references [1]–[29].

In this paper, we shall also adopt the statistical mechanical viewpoint on channel coding. We focus on the classical random code ensemble (RCE) for communicating over a discrete memoryless channel (DMC), in the same setting as described in [19, Chap. 6] and [24], which in a nutshell, is as follows: Consider a DMC, $P(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^n p(y_i|x_i)$, fed by an input n–vector that belongs to a codebook $\mathcal{C} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\}$, $M = e^{nR}$, with uniform priors, where R is the coding rate in nats per channel use. The induced posterior, for $\boldsymbol{x} \in \mathcal{C}$, is then:

$$\begin{aligned} P(\boldsymbol{x}|\boldsymbol{y}) &= \frac{P(\boldsymbol{y}|\boldsymbol{x})}{\sum_{\boldsymbol{x}' \in \mathcal{C}} P(\boldsymbol{y}|\boldsymbol{x}')} \\ &= \frac{e^{-\ln[1/P(\boldsymbol{y}|\boldsymbol{x})]}}{\sum_{\boldsymbol{x}' \in \mathcal{C}} e^{-\ln[1/P(\boldsymbol{y}|\boldsymbol{x}')]}}. \end{aligned} \tag{1}$$

Here, the second line is written in a form that resembles the Boltzmann distribution of statistical physics, according to which the probability of a certain 'state' (or 'configuration') of the system, designated by $\boldsymbol{x}$, is given by

$$P(\boldsymbol{x}) = \frac{e^{-\beta \mathcal{E}(\boldsymbol{x})}}{Z(\beta)} \tag{2}$$

where $\beta = 1/(kT)$ is the inverse temperature, k is Boltzmann's constant,[1] T is temperature, $\mathcal{E}(\boldsymbol{x})$ is the energy associated with $\boldsymbol{x}$, and $Z(\beta) = \sum_{\boldsymbol{x}} e^{-\beta \mathcal{E}(\boldsymbol{x})}$ is the *partition function*. In our case, of course, $\beta = 1$ and the energy function (which depends on the given $\boldsymbol{y}$) is $\mathcal{E}(\boldsymbol{x}) = \ln[1/P(\boldsymbol{y}|\boldsymbol{x})]$. But this analogy with the Boltzmann distribution (2) naturally suggests (cf. e.g., [19]) to consider,

[1] Here we will adopt the convention, customarily used in many papers and books, of redefining 'temperature' according to $T \leftarrow kT$, that is, in units of energy, and then $\beta \stackrel{\triangle}{=} 1/T$.

more generally, the posterior distribution parametrized by β, that is

$$P_\beta(\boldsymbol{x}|\boldsymbol{y}) = \frac{P^\beta(\boldsymbol{y}|\boldsymbol{x})}{\sum_{\boldsymbol{x}'\in\mathcal{C}} P^\beta(\boldsymbol{y}|\boldsymbol{x}')} = \frac{e^{-\beta\ln[1/P(\boldsymbol{y}|\boldsymbol{x})]}}{\sum_{\boldsymbol{x}'\in\mathcal{C}} e^{-\beta\ln[1/P(\boldsymbol{y}|\boldsymbol{x}')]}} \stackrel{\Delta}{=} \frac{e^{-\beta\ln[1/P(\boldsymbol{y}|\boldsymbol{x})]}}{Z(\beta|\boldsymbol{y})}. \tag{3}$$

There are a few motivations for introducing the temperature parameter in (3). First, it allows a degree of freedom in case there is some uncertainty regarding the channel noise level (small β corresponds to high noise level). Second, it is inspired by the ideas behind simulated annealing techniques: by sampling from P_β while gradually increasing β (cooling the system), the minima of the energy function (ground states) can be found. Third, by applying symbolwise MAP decoding, i.e., decoding the ℓ–th symbol of $\boldsymbol{x}$ as $\arg\max_a P_\beta(x_\ell = a|\boldsymbol{y})$, where

$$P_\beta(x_\ell = a|\boldsymbol{y}) = \sum_{\boldsymbol{x}\in\mathcal{C}:\ x_\ell=a} P_\beta(\boldsymbol{x}|\boldsymbol{y}),$$

we obtain a family of *finite–temperature decoders* (originally proposed by Ruján [26]; see also [4], [19, Section 6.3.3],[29],[27]) parametrized by β, where $\beta = 1$ corresponds to minimum symbol error probability (with respect to the true channel) and $\beta \to \infty$ corresponds to minimum block error probability. Finally, and this is the motivation that drives the research reported in this paper: the corresponding partition function, $Z(\beta|\boldsymbol{y})$, namely, the sum of (conditional) probabilities raised to some power β, is an expression frequently encountered in Rényi information measures as well as in the analysis of random coding exponents using Gallager's techniques. Since the partition function plays a key role in statistical mechanics, as many physical quantities can be derived from it, then it is natural to ask if it can also be used to gain some insights regarding the behavior of random codes at various temperatures and coding rates. The main contribution of this paper is in exploring this direction.

To sharpen the last point a little further, it is noted that when one considers the random coding regime, as we do in this paper, then even if $\boldsymbol{y}$ is given, the energy levels pertaining to the Boltzmann distribution (3) are themselves random variables since they depend on the randomly chosen codevectors. As explained in [19], this then falls under the umbrella of the so called *random energy model* (REM) in statistical physics, which was invented by Derrida [30] with the motivation

to capture disorder in spin glass systems. The interesting fact about the REM is that it is typically subjected to *phase transitions*, and then so is the model (3) for random codes.

More specifically, as described in [19, Chap. 6], [25], and as will be briefly reviewed in the next section, the partition function pertaining to finite–temperature decoding of a (typical) randomly chosen code is known to have three types of behavior, corresponding to three phases in the plane of rate vs. temperature: the *ferromagnetic phase*, corresponding to correct decoding, the *paramagnetic phase*, of complete disorder, which is dominated by exponentially many incorrect codewords, and the *glassy phase* (or the condensed phase), where the system is frozen at minimum energy and dominated by subexponentially many incorrect codewords. We show that the statistical physics associated with the two latter phases are intimately related to random coding exponents. In particular, the exponent associated with the probability of correct decoding at rates above capacity is directly related to the free energy in the glassy phase, and the exponent associated with probability of error (the error exponent) at rates below capacity, is strongly related to the free energy in the paramagnetic phase. In fact, we derive alternative expressions of these exponents in terms of the corresponding free energies, and make an attempt to obtain some insights from these expressions.

An additional interesting byproduct of the statistical mechanical point of view that we adopt in this work, is that it suggests a more refined analysis technique, as an alternative to the customary use of Jensen's inequality, for which it is clear that the resulting expressions are exponentially tight, and not just bounds. Another way to look at this is to observe that the analysis technique, inspired by statistical mechanical point of view, provides us with insights with regard to the conditions under which Jensen's inequality provides a tight bound in this context. We believe that this technique may be useful in other applications as well. We shall elaborate more on this in the sequel.

As a side result, we also compare the phase diagram associated with a certain universal decoder (namely, the minimum conditional entropy universal decoder) for discrete memoryless channels, to that of the finite–temperature decoder that is aware of the channel statistics, and show that in spite of the fact that this universal decoder is asymptotically optimum, in the sense of attaining optimum random coding error exponents [31], its phase diagram is substantially different.

The outline of the remaining part of this paper is as follows. In Section 3, we provide some background, which mostly follows the presentation in [19] (with a few missing details filled in), but

will be useful here to keep this paper self contained. Section 3 also includes a subsection with the phase diagram for universal decoding, as described in the previous paragraph. In Section 4, we derive the alternative formula for the exponent of correct decoding above capacity, and in Section 5, we do the same regarding the random coding exponent at rates below capacity.

2 Notation Conventions, Background and Preliminaries

2.1 Notation Conventions

Throughout this paper, scalar random variables (RV's) will be denoted by capital letters, like X and Y, their sample values will be denoted by the respective lower case letters, and their alphabets will be denoted by the respective calligraphic letters. A similar convention will apply to random vectors and their sample values, which will be denoted with the same symbols in the boldface font. Thus, for example, $\boldsymbol{X}$ will denote a random n-vector $(X_1, \ldots, X_n)$, and $\boldsymbol{x} = (x_1, ..., x_n)$ is a specific vector value in $\mathcal{X}^n$, the n-th Cartesian power of $\mathcal{X}$.

Sources and channels will be denoted generically by the letters P and Q. Specific letter probabilities corresponding to a source Q will be denoted by the corresponding lower case letters, e.g., $q(x)$ is the probability of a letter $x \in \mathcal{X}$. A similar convention will be applied to the channel P and the corresponding transition probabilities, $p(y|x)$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$. The expectation operator will be denoted by $\boldsymbol{E}\{\cdot\}$.

The empirical distribution pertaining to a vector $\boldsymbol{x} \in \mathcal{X}^n$ will be denoted by $\hat{P}_{\boldsymbol{x}}$. In other words, $\hat{P}_{\boldsymbol{x}} = \{\hat{p}_{\boldsymbol{x}}(a),\ a \in \mathcal{X}\}$, where $p_{\boldsymbol{x}}(a) = n_{\boldsymbol{x}}(a)/n$, $n_{\boldsymbol{x}}(a)$ being the number of occurrences of the letter a in $\boldsymbol{x}$. Similar conventions will apply to empirical joint distributions of pairs of letters, $(a, b) \in \mathcal{X} \times \mathcal{Y}$, extracted from the corresponding pairs of vectors $(\boldsymbol{x}, \boldsymbol{y})$, that is, the joint empirical distribution $\hat{P}_{\boldsymbol{xy}}$ is the vector of relative frequencies of joint occurrences of $x_i = a$ and $y_i = b$, $i = 1, \ldots, n$. Similarly, $\hat{p}_{\boldsymbol{x}|\boldsymbol{y}}(a|b) = \hat{p}_{\boldsymbol{xy}}(a, b)/\hat{p}_{\boldsymbol{y}}(b)$ will denote the empirical conditional probability of $X = a$ given $Y = b$ (with convention that $0/0 = 0$), and $\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}$ will denote $\{\hat{p}_{\boldsymbol{x}|\boldsymbol{y}}(a|b),\ a \in \mathcal{X},\ b \in \mathcal{Y}\}$. The expectation w.r.t. the empirical distribution of $(\boldsymbol{x}, \boldsymbol{y})$ will be denoted by $\hat{\boldsymbol{E}}_{\boldsymbol{xy}}\{\cdot\}$, i.e., for a given function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, we define $\hat{\boldsymbol{E}}_{\boldsymbol{xy}}\{f(X, Y)\}$ as $\sum_{(a,b)\in\mathcal{X}\times\mathcal{Y}} \hat{p}_{\boldsymbol{xy}}(a, b) f(a, b)$, where in this notation, X and Y are understood to be random variables jointly distributed according to $\hat{P}_{\boldsymbol{xy}}$.

The cardinality of a finite set $\mathcal{A}$ will be denoted by $|\mathcal{A}|$. For two positive sequences $\{a_n\}$ and $\{b_n\}$, the notation $a_n \doteq b_n$ means that a_n and b_n are asymptotically of the same exponential order, that is, $\lim_{n\to\infty} \frac{1}{n} \ln \frac{a_n}{b_n} = 0$. Information theoretic quantities like entropies and mutual informations will be denoted following the usual conventions of the Information Theory literature. When we wish to make it clear that such an information theoretic quantity is induced by a certain probability distribution, say Q, we use this probability distribution as a subscript, e.g., $I_Q(X;Y)$, $H_Q(X|Y)$, etc. When the underlying probability distribution is an empirical distribution, we will subscript it by the sequences(s) from which the empirical distribution is extracted, and we will use hats, e.g., $\hat{I}_{\boldsymbol{x}\boldsymbol{y}}(X;Y)$, $\hat{H}_{\boldsymbol{x}\boldsymbol{y}}(X|Y)$.

2.2 Background and Preliminaries

Consider a DMC with a finite input alphabet $\mathcal{X}$ and a finite output alphabet $\mathcal{Y}$, which when fed by an input vector $\boldsymbol{x} \in \mathcal{X}^n$, it generates an output vector $\boldsymbol{y} \in \mathcal{Y}^n$ distributed according to

$$P(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} p(y_i|x_i),$$

where $\{p(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$ are given single–letter transition probabilities. Let

$$\mathcal{C} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\} \subseteq \mathcal{X}^n$$

be a codebook of $M = e^{nR}$ codewords, where R is the coding rate (in nats per channel use). Next consider the posterior distribution (3) and the corresponding partition function

$$Z(\beta|\boldsymbol{y}) = \sum_{\boldsymbol{x}\in\mathcal{C}} P^{\beta}(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\boldsymbol{x}\in\mathcal{C}} e^{-\beta d(\boldsymbol{x},\boldsymbol{y})}, \tag{4}$$

where $d(\boldsymbol{x},\boldsymbol{y}) = -\ln P(\boldsymbol{y}|\boldsymbol{x}) = -\sum_{i=1}^{n} \ln P(y_i|x_i)$. We shall think of $Z(\beta|\boldsymbol{y})$ as the sum of two contributions, the first is $Z_c(\beta|\boldsymbol{y}) = e^{-\beta d(\boldsymbol{x}_0,\boldsymbol{y})}$, pertaining to the correct codeword $\boldsymbol{x}_0$ (that was actually transmitted across the channel), and the second is associated with the remaining (incorrect) codewords,

$$Z_e(\beta|\boldsymbol{y}) = \sum_{\boldsymbol{x}\in\mathcal{C}-\{\boldsymbol{x}_0\}} e^{-\beta d(\boldsymbol{x},\boldsymbol{y})}.$$

Let us focus on $Z_e(\beta|\boldsymbol{y})$ first. As mentioned in the Introduction, when the codebook $\mathcal{C}$ is selected at random, this is a disordered system in the framework of the REM, which exhibits phase transitions.

To describe these phase transitions, it is instructive to begin with the relatively simple special case of the binary symmetric channel (BSC), as we do in Subsection 2.2.1, and then extend the scope to general DMC's, as in Subsection 2.2.2.[2] Finally, Subsection 2.2.3 (which is not included in [19]) is about a phase diagram pertaining to universal decoding (cf. second to the last paragraph of the Introduction). This subsection can be skipped without loss of continuity.

2.2.1 The Binary Symmetric Channel

For the BSC with a crossover parameter p, we have $P(\boldsymbol{y}|\boldsymbol{x}) = p^{d_H(\boldsymbol{x},\boldsymbol{y})}(1-p)^{n-d_H(\boldsymbol{x},\boldsymbol{y})}$, where $d_H(\boldsymbol{x},\boldsymbol{y})$ is the Hamming distance between $\boldsymbol{x}$ and $\boldsymbol{y}$. Defining $B = \ln\frac{1-p}{p}$, we then have $P(\boldsymbol{y}|\boldsymbol{x}) = (1-p)^n e^{-Bd_H(\boldsymbol{x},\boldsymbol{y})}$, and so

$$\begin{aligned} Z_e(\beta|\boldsymbol{y}) &= \sum_{\boldsymbol{x}\in\mathcal{C}} P^\beta(\boldsymbol{y}|\boldsymbol{x}) \\ &= (1-p)^{\beta n}\sum_{\boldsymbol{x}\in\mathcal{C}} e^{-\beta B d_H(\boldsymbol{x},\boldsymbol{y})} \\ &= (1-p)^{\beta n}\sum_{d=0}^{n} N_{\boldsymbol{y}}(d)e^{-\beta B d}, \end{aligned} \tag{5}$$

where $N_{\boldsymbol{y}}(d)$ is the number of incorrect codewords at Hamming distance d from $\boldsymbol{y}$. As argued in [19], when the codewords are chosen independently at random (say, by fair coin tossing), $\{N_{\boldsymbol{y}}(d)\}$ concentrate very rapidly,[3] as $n\to\infty$, about their expectations:

$$\boldsymbol{E}\{N_{\boldsymbol{y}}(\delta n)\} \doteq e^{n[R-\ln 2+h(\delta)]}, \quad 0\le\delta\le 1 \tag{6}$$

where $h(\delta) \triangleq -\delta\ln\delta-(1-\delta)\ln(1-\delta)$. Defining the normalized Gilbert–Varshamov (GV) distance, $\delta_{GV}(R)$, as the solution, δ, to the equation $h(\delta)=\ln 2 - R$, it is apparent that for $\delta<\delta_{GV}(R)$ and $\delta > 1-\delta_{GV}(R)$, $\boldsymbol{E}\{N_{\boldsymbol{y}}(\delta n)\}$ has a negative exponent, and thus typically, these distances are not populated by codewords. Therefore, for a typical random code,

$$\begin{aligned} Z_e(\beta|\boldsymbol{y}) &\doteq (1-p)^{\beta n}\cdot e^{n(R-\ln 2)}\int_{\delta_{GV})}^{1-\delta_{GV}(R)} \mathrm{d}\delta\cdot e^{nh(\delta)}\cdot e^{-\beta B\delta} \\ &\doteq (1-p)^{\beta n}\cdot e^{n(R-\ln 2)}\exp\left\{n\cdot \max_{\delta\in[\delta_{GV}(R),1-\delta_{GV}(R)]}[h(\delta)-\beta B\delta]\right\} \\ &\triangleq e^{-n\beta F_e(\beta)} \end{aligned} \tag{7}$$

[2] This extension to general DMC's in outlined in [19, Chap. 6], but here we provide some more details.

[3] Note that $N_{\boldsymbol{y}}(d) = \sum_{i=1}^{e^{nR}} 1\{d_H(\boldsymbol{x}_i,\boldsymbol{y}) = d\}$, i.e., it is the sum of exponentially many i.i.d. (given $\boldsymbol{y}$) random variables, and so, its large deviations behavior is exponential in e^{nR}, which is double–exponential in n (see also Appendix, Subsection A.2.).

where $F_e(\beta)$ is the *free energy density* associated with the incorrect codewords, which is given by

$$F_e(\beta) = \begin{cases} \delta_{GV}(R)\ln\frac{1}{p} + (1-\delta_{GV}(R))\ln\frac{1}{1-p} & p_\beta \le \delta_{GV}(R) \\ \frac{1}{\beta}[\ln 2 - R - \ln(p^\beta + (1-p)^\beta)] & p_\beta > \delta_{GV}(R) \end{cases} \tag{8}$$

where

$$p_\beta = \frac{p^\beta}{p^\beta + (1-p)^\beta},$$

and where the distinction between the two different expressions is due to the constraint $\delta \in [\delta_{GV}(R), 1-\delta_{GV}(R)]$, which becomes active (i.e., achieved with equality) when $p_\beta \le \delta_{GV}(R)$. We observe then that when p and R are held fixed, and β varies, the above expression exhibits a phase transition at temperature $T_c(R) = 1/\beta_c(R)$ for which $p_\beta = \delta_{GV}(R)$, i.e.,

$$\beta_c(R) = \frac{\ln[(1-\delta_{GV}(R))/\delta_{GV}(R)]}{\ln[(1-p)/p]}.$$

For $\beta > \beta_c$ (low temperature), the free energy density $F_e(\beta) = \delta_{GV}(R)\ln\frac{1}{p} + (1-\delta_{GV}(R))\ln\frac{1}{1-p}$ is independent of β hence the entropy (which is related to the derivative of $F_e(\beta)$ w.r.t. β) vanishes, and the system is frozen in the sense that the thermodynamics are dominated by a subexponential number of configurations of the minimum energy which is $n\delta_{GV}(R)$. This phase is referred to as *condensed phase* or *glassy phase*, and henceforth we denote

$$F_g \triangleq \delta_{GV}(R)\ln\frac{1}{p} + (1-\delta_{GV}(R))\ln\frac{1}{1-p}.$$

For $\beta > \beta_c$, the thermodynamics are dominated by an exponential number of states at distance np_β, which is larger than $n\delta_{GV}(R)$, and the entropy is strictly positive. This is called the *paramagnetic phase* and henceforth we denote

$$F_p(\beta) \triangleq \frac{1}{\beta}[\ln 2 - R - \ln(p^\beta + (1-p)^\beta)].$$

When the contribution of $Z_c(\beta) = e^{-n\beta F_c}$ is taken into account, and we consider the total partition function $Z(\beta)$, the situation changes: Since $d_H(\boldsymbol{x}_0, \boldsymbol{y})$ is typically about the level of np, and thus the corresponding free energy density is $F_c = h(p)$, we have yet another phase referred to as the *ordered phase* or the *ferromagnetic phase*. This phase exists whenever $Z(\beta)$ is dominated by $Z_c(\beta)$, i.e., $F_c = h(p) < F_e(\beta)$. For $\beta > \beta_c$, this is the case whenever $p < \delta_{GV}(R)$, or equivalently, $R < \ln 2 - h(p) \triangleq C$, where C is the capacity of the BSC. For $\beta < \beta_c$ the boundary between the

ferromagnetic phase and the paramagnetic phase is given by the solution $\beta_0(R) = 1/T_0(R)$ to the equation

$$\beta h(p) = \ln 2 - R - \ln[p^\beta + (1-p)^\beta]. \tag{9}$$

To summarize, while there are only two phases (glassy and paramagnetic) pertaining to $Z_e(\beta)$, there is a third, additional phase (ferromagnetic) associated with $Z_c(\beta)$. In the ferromagnetic phase, the system is dominated by one state corresponding to the correct codeword. Thus, similarly as in the glassy phase, the entropy of the ferromagnetic phase is zero. The boundaries between the three phases in the plane defined by R and $T = 1/\beta$, are as follows (see Fig. 1): The ferro–glassy boundary is the straight line $R = C$, the glassy–paramagnetic boundary is the curve $T = T_c(R)$, and the and the ferro–paramagnetic boundary $T = T_0(R)$ is given by eq. (9). The triple point where all boundaries intersect is the point $(R, T) = (C, 1)$.

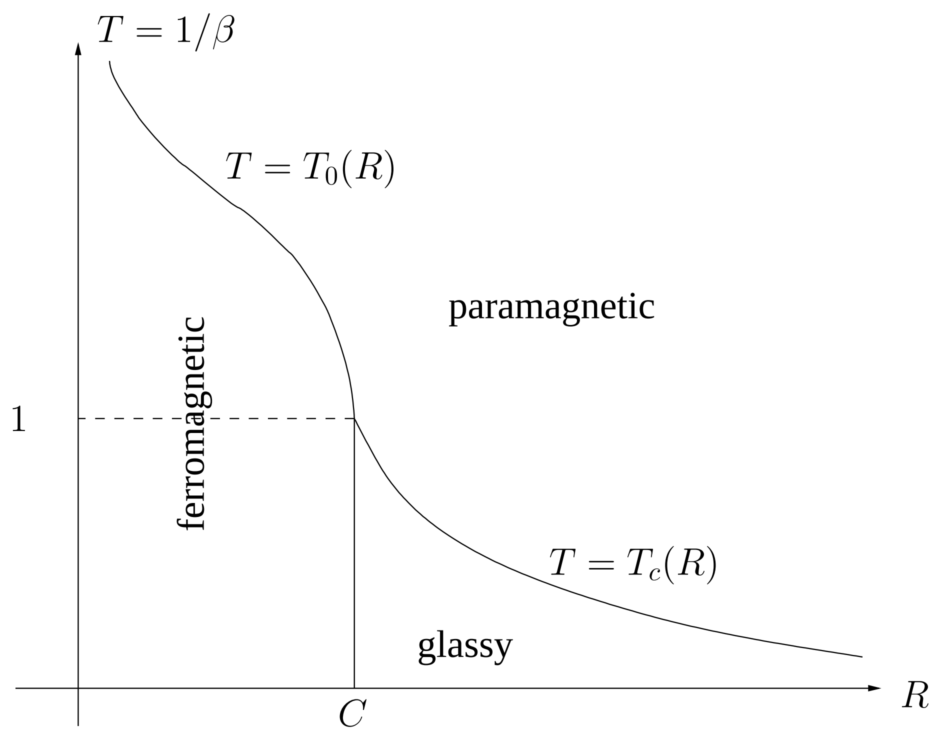

Figure 1: Phase diagram of the finite–temperature MAP decoder.

In spite of the fact that in the glassy phase there are only few configurations that dominate the behavior, it is no different from the paramagnetic phase in terms of the typical ranking of the likelihood of the correct codeword among all codewords: In both phases, the typical location of

the correct codeword in the list of descending likelihoods, $\{P(\boldsymbol{y}|\boldsymbol{x}_i)\}$, is about $2^{n(R-C)}$ $(R > C)$. Although the glassy phase exhibits less uncertainty, or equivalently, more certainty, (sublinear conditional entropy given $\boldsymbol{y}$ about the channel input), this relative certainty is misleading because the posterior probability mass is captured mostly by incorrect codewords. In this sense, the glassy phase is even more problematic than the paramagnetic one: Since the certainty is fictitious, it is more difficult to detect errors.

2.2.2 Extension to General DMC's

The extension to general DMC's is essentially quite straightforward. Consider a DMC parametrized by $\{P(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$. For the sake of simplicity, let us consider the uniform random coding distribution[4] according to which each codeword is selected independently at random with probability distribution $Q(\boldsymbol{x}) = 1/|\mathcal{X}|^n$ for all $\boldsymbol{x} \in \mathcal{X}^n$. For a given channel output vector $\boldsymbol{y}$, the probability of selecting a random codeword $\boldsymbol{x}$ whose conditional empirical distribution with $\boldsymbol{y}$ is $\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}$ is of the exponential order of $e^{-n[\ln|\mathcal{X}|-\hat{H}_{\boldsymbol{xy}}(X|Y)]}$, [32], thus the expected number of codewords with this conditional distribution is exponentially

$$\boldsymbol{E}\{N_{\boldsymbol{y}}(\hat{P}_{\boldsymbol{x}|\boldsymbol{y}})\} \doteq e^{n[R-\ln|\mathcal{X}|+\hat{H}_{\boldsymbol{xy}}(X|Y)]}.$$

In analogy to the explanation provided in the previous subsection (and in [19]), in the context of the BSC, those conditional distributions $\{\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}\}$ for which the exponent on the right–hand side is negative, are typically not populated. Thus, for a typical random code

$$\begin{aligned}
Z_e(\beta|\boldsymbol{y}) &= \sum_{\boldsymbol{x}\in\mathcal{C}-\{\boldsymbol{x}_0\}} P^{\beta}(\boldsymbol{y}|\boldsymbol{x}) \\
&= \sum_{\boldsymbol{x}\in\mathcal{C}-\{\boldsymbol{x}_0\}} e^{-\beta\ln 1/P(\boldsymbol{y}|\boldsymbol{x})} \\
&= \sum_{\{\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}\}} N_{\boldsymbol{y}}(\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}) \cdot \exp\{-\beta\hat{\boldsymbol{E}}_{\boldsymbol{xy}}\ln[1/P(Y|X)]\} \\
&\doteq \exp\left\{n\left(R-\ln|\mathcal{X}| + \max_{Q_{X|Y}:\ H_Q(X|Y)\geq\ln|\mathcal{X}|-R}[H_Q(X|Y)-\beta\boldsymbol{E}_Q\{\ln[1/P(Y|X]\}]\right)\right\} \\
&\stackrel{\Delta}{=} e^{-n\beta F_e(\beta;Y)},
\end{aligned} \tag{10}$$

[4]Other random coding distributions can be used as well, but will lead to somewhat more complicated expressions, which we prefer to avoid in this description.

where Y designates a RV distributed according to the empirical distribution $\hat{P}_{\boldsymbol{y}}$ of $\boldsymbol{y}$.

A word on notation is now in order: here and throughtout the sequel, we adopt the common abuse of notation, customarily used in the Information Theory literature, that when a RV appears as an argument or a subscript of a certain function, this means that it is actually a functional of its distribution, not a function of the value of the random variable itself. Whenever we wish to emphasize the dependence of this quantity on the empirical distribution $\hat{P}_{\boldsymbol{y}}$, we will replace Y by $\hat{P}_{\boldsymbol{y}}$ or simply by $\boldsymbol{y}$ itself, provided that the context does not leave room for ambiguity. Similar comments will apply to other quantities to be defined throughout this subsection and in the sequel.[5] For some of these quantites, we will not denote the dependence on the distribution of Y explicitly, in order to avoid cumbersome notation, but it will be made clear that they do depend on it in general.

Consider now the expression

$$J_Y(\beta, R) \stackrel{\Delta}{=} \max_{Q_{X|Y}: H_Q(X|Y) \geq \ln|\mathcal{X}| - R} \left[H_Q(X|Y) - \beta \boldsymbol{E}_Q\{d(X,Y)\} \right],$$

where $d(x,y) \stackrel{\Delta}{=} -\ln p(y|x)$.

First, it is easy to prove (see Appendix, Subsection A.1) that for fixed β and $\boldsymbol{y}$, the function $J_Y(\beta, R)$ is concave in R. This means that the inequality constraint $H_Q(X|Y) \geq \ln|\mathcal{X}| - R$ is met with equality as long as $R \leq R_Y(\beta)$, where $R_Y(\beta) = \ln|\mathcal{X}| - H_{Q_\beta}(X|Y)$ with Q_β being the achiever of

$$J_Y(\beta, \ln|\mathcal{X}|) = \max_{Q_{X|Y}} [H_Q(X|Y) - \beta \boldsymbol{E}_Q\{d(X,Y)\}],$$

that is,

$$Q_\beta(x|y) = \frac{e^{-\beta d(x,y)}}{\sum_{x' \in \mathcal{X}} e^{-\beta d(x',y)}} = \frac{P^\beta(y|x)}{\sum_{x' \in \mathcal{X}} P^\beta(y|x')}.$$

We will also use the notation

$$D_Y(\beta) = \boldsymbol{E}_{Q_\beta}\{d(X,Y)\}$$

and

$$H_Y(\beta) = H_{Q_\beta}(X|Y),$$

[5] In Subsection 2.2.1, this issue did not arise since all relevant quantities happened to be independent of $\hat{P}_{\boldsymbol{y}}$, due to the symmetry of the BSC.

thus $R_Y(\beta) = \ln|X| - H_Y(\beta)$. Let

$$\beta_c(R) \stackrel{\Delta}{=} \inf\{\beta : \; R_Y(\beta) \geq R\} = \inf\{\beta : \; H_Y(\beta) \leq \ln|\mathcal{X}| - R\}.$$

Obviously, $\beta_c(R)$ increases with R, or equivalently, $T_c(R) = 1/\beta_c(R)$ is decreasing with R ($T_c(\ln|\mathcal{X}|) = 0$). This forms the boundary curve between the glassy and the paramagnetic phases. Note that when $R = I(X;Y)$, the mutual information induced by the uniform distribution on $\mathcal{X}$ and by $P(y|x)$, then $\beta_c(R) = 1$. Thus, $(I(X;Y), 1)$ is a point on the curve $T = T_c(R)$.

For $R \leq R_Y(\beta)$, or equivalently, $\beta \geq \beta_c(R)$, the constraint $H_Q(X;Y) \geq \ln|\mathcal{X}| - R$ is attained with equality. Thus, in this range of low rates,

$$\begin{aligned} J_Y(\beta, R) &= \max_{\{Q_{X|Y}: H_Q(X|Y) = \ln|\mathcal{X}| - R\}} [\ln|\mathcal{X}| - R - \beta \boldsymbol{E}_Q\{d(X,Y)\}] \\ &= \ln|\mathcal{X}| - R - \beta \cdot \min_{\{Q_{X|Y}: H_Q(X|Y) = \ln|\mathcal{X}| - R\}} \boldsymbol{E}_Q\{d(X,Y)\} \\ &= \ln|\mathcal{X}| - R - \beta D_Y(\beta_R) \end{aligned} \tag{11}$$

where β_R is the solution to the equation $H_Y(\beta) = \ln|\mathcal{X}| - R$. We will also use the notation $\delta_Y(R) = D_Y(\beta_R)$.[6] It follows then that $F_e(\beta, Y) = F_g(Y) = \delta_Y(R)$, which is the glassy phase.

For $R > R_Y(\beta)$,

$$J_Y(\beta, R) = J_Y(\beta, \ln|\mathcal{X}|) = \max_{Q_{X|Y}} [H_Q(X|Y) - \beta \boldsymbol{E}_Q\{d(X,Y)\}] = H_Y(\beta) - \beta D_Y(\beta)$$

Thus, for $\beta < \beta_c(R)$,

$$F_e(\beta, Y) = F_p(\beta, Y) = D_Y(\beta) + \frac{\ln|\mathcal{X}| - R - H_Y(\beta)}{\beta},$$

which is the paramagnetic phase. It should be pointed out that for a general decoding metric $d(x,y)$ (not necessarily ML matched to the channel), the boundary between the paramagnetic and the glassy phases depends only on the random coding distribution and this decoding metric $d(x,y)$, not on the channel itself (cf. Subsection 2.2.3). The boundaries with the ferromagnetic phase are the ones that depend on the channel.

In the ordered (ferromagnetic) phase, the free energy density is given by $F(\beta) = H(Y|X)$, where X is uniform and Y given X is distributed according to the channel. As long as $R < I(X;Y)$,

[6] The quantity $\delta_Y(R)$ is the generalization of the GV distance that was defined in Subsection 2.2.1. for the BSC.

we have $H(Y|X) < \delta_Y(R)$. In fact, the line connecting the points $(R = I(X;Y), T = 1)$ and $(R = I(X;Y), T = 0)$ forms the boundary between the ordered ferromagnetic phase and the glassy phase.

For $R < I(X;Y)$, the boundary between the ferromagnetic and paramagnetic phases is given by the solution $\beta_0(R)$ (or $T_0(R) = 1/\beta_0(R)$) to the equation

$$\beta H(Y|X) = \beta D_Y(\beta) + \ln|\mathcal{X}| - R - H_Y(\beta),$$

which is above the curve $T = T_c(R)$ for $R < I(X;Y)$. It should be emphasized that $\beta_c(R)$, $\beta_0(R)$, and β_R all depend on the (distribution of the) RV Y, namely, the empirical distribution of $\boldsymbol{y}$.

2.2.3 Phase Diagram for Universal Decoding

It is instructive to compare the phase diagram of finite–temperature MAP decoding to those of finite–temperature universal decoders. One simple example of a universal decoder for which it is especially easy to derive the phase diagram is the minimum conditional entropy decoder [31], which given $\boldsymbol{y}$, selects the codeword $\boldsymbol{x}_m$ for which $\hat{H}_{\boldsymbol{x}_m\boldsymbol{y}}(X|Y)$ is minimum.[7] It is well known that this universal decoder is asymptotically optimum in the random coding sense, in that it achieves the same random coding error exponent as the ML decoder, provided that the random coding distribution is uniform over $\mathcal{X}^n$.

The partition function corresponding to this universal decoder is the same as before, except that $\hat{\boldsymbol{E}}_{\boldsymbol{x}\boldsymbol{y}}\{d(X,Y)\}$ is replaced by $\hat{H}_{\boldsymbol{x}\boldsymbol{y}}(X|Y)$. In this case,

$$\begin{aligned} Z_e(\beta|\boldsymbol{y}) &= \sum_{\{\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}\}} N_{\boldsymbol{y}}(\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}) \cdot e^{-\beta \hat{H}_{\boldsymbol{x}\boldsymbol{y}}(X|Y)} \\ &\doteq \exp\left\{n\left(R - \ln|\mathcal{X}| + \max_{Q_{X|Y}:\ H_Q(X|Y)\geq \ln|\mathcal{X}|-R}[(1-\beta)H_Q(X|Y)]\right)\right\} \\ &\stackrel{\Delta}{=} e^{-n\beta F_e(\beta,Y)} \end{aligned} \qquad (12)$$

Now, it is easy to see how phase transitions behave (see Fig. 2): If $\beta < 1$, then the maximum is $\ln|\mathcal{X}|$ and we get

$$Z_e(\beta) \doteq e^{n[R-\beta\ln|\mathcal{X}|]},$$

[7] This is a variant of the well–known maximum mutual information (MMI) decoder. In the case of constant composition codes, these two decoders are identical.

thus $F_e(\beta, Y) = F_p(\beta) = \ln|\mathcal{X}| - R/\beta$. If $\beta > 1$, we get

$$Z_e(\beta) \doteq e^{-n\beta[\ln|\mathcal{X}|-R]},$$

thus, $F_e(\beta, Y) = F_g = \ln|\mathcal{X}| - R$. Therefore, the boundary between the two phases is the horizontal line $T_c = 1/\beta_c = 1$ (independently of R). This means that the glassy region here is larger than in ML decoding for $R > C$. The boundary between the ferromagnetic and the glassy phases continues to be $R = I(X;Y)$ as before. The ferromagnetic–paramagnetic boundary is now $H(X|Y) = \ln|\mathcal{X}| - R/\beta$, or, equivalently, $T = 1/\beta = I(X;Y)/R$, which is below the ferromagnetic–paramagnetic boundary of the MAP decoder. This can easily be shown by setting $R = \beta I(X;Y)$ (which is this boundary) in the r.h.s. of the equation defining $T_0(R)$ and showing that the resulting expression is larger than $\beta H(Y|X)$ (for $\beta \leq 1$), which is the l.h.s. of this equation (thus, we are still in the ferromagnetic phase of MAP decoding): Specifically, the l.h.s. of the equation defining $T_0(R)$ is:

$$\begin{aligned}
& \beta D_Y(\beta) + \ln|\mathcal{X}| - R - H_Y(\beta) \\
= \; & \beta D_Y(\beta) + \ln|\mathcal{X}| - \beta I(X;Y) - H_Y(\beta) \\
= \; & \beta H(Y|X) + \beta \boldsymbol{E}_{Q_\beta} \ln \frac{1}{P(Y|X)} + \ln|\mathcal{X}| - \beta H(Y) - H_{Q_\beta}(X|Y) \\
= \; & \beta H(Y|X) + \beta \boldsymbol{E}_{Q_\beta} \ln \frac{1}{P(Y|X)} - \beta H(Y) + I_{Q_\beta}(X;Y) \\
\geq \; & \beta H(Y|X) + \beta H_{Q_\beta}(Y|X) - \beta H(Y) + I_{Q_\beta}(X;Y) \\
\geq \; & \beta H(Y|X) - \beta I_{Q_\beta}(X;Y) + I_{Q_\beta}(X;Y) \\
\geq \; & \beta H(Y|X)
\end{aligned} \tag{13}$$

where the first equality is since $R = \beta I(X;Y)$ on the boundary, and the last equality is since $\beta \leq 1$. Thus, although this decoder achieves the optimum random coding error exponent, it has a phase diagram which is worse than that of MAP decoding, as the ferromagnetic region is smaller and the glassy region is larger.

3 The Correct Decoding Exponent

We now proceed to establish relationships between the phase diagram of a random code, decoded by a finite temperature MAP decoder, and the exponent of correct decoding at rates above capacity,

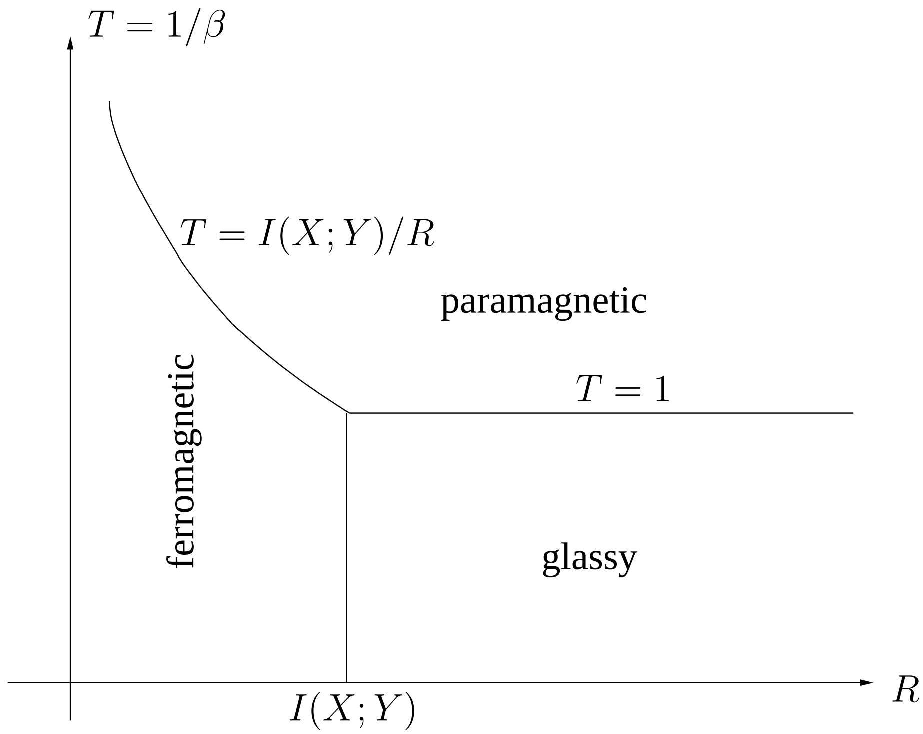

Figure 2: Phase diagram for universal decoding.

or to be more precise, rates above $I(X;Y)$, the mutual information induced by the uniform input distribution and the channel.

Arimoto [33] begins the derivation of his bound on the probability of correct decoding by using the inequality

$$P_c = \frac{1}{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\max_{1\le i\le M} P(\boldsymbol{y}|\boldsymbol{x}_i) \le \frac{1}{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\left[\sum_{i=1}^{M} P^{\beta}(\boldsymbol{y}|\boldsymbol{x}_i)\right]^{1/\beta}, \quad \beta > 0 \tag{14}$$

which becomes tight when $\beta \to \infty$. We will also use this inequality, but we shall proceed somewhat differently than Arimoto. First, observe that for a randomly selected code, where the average probability of correct decoding is upper bounded by

$$\bar{P}_c \le \frac{1}{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\boldsymbol{E}\left\{\left[\sum_{i=1}^{M} P^{\beta}(\boldsymbol{y}|\boldsymbol{x}_i)\right]^{1/\beta}\right\}, \tag{15}$$

the expression in the square brackets is exactly $Z_e(\beta)$ (just with M codewords instead of $M-1$), because the interpretation of this expression, is that the codewords are drawn under Q regardless

of $\boldsymbol{y}$. Since we are interested in $\beta \to \infty$ (in addition to the assumption that $R > I(X;Y)$), then we are actually carrying out this calculation in the *glassy* regime.

The above upper bound to $\bar{P}_c$ can be also written as:

$$\bar{P}_c \leq \frac{1}{M} \sum_{\boldsymbol{y} \in \mathcal{Y}^n} \boldsymbol{E} \left\{ \left[\sum_{d \in \mathcal{D}_n} N_{\boldsymbol{y}}(d) \cdot e^{-\beta d} \right]^{1/\beta} \right\}, \tag{16}$$

where here $N_{\boldsymbol{y}}(d)$ denotes the number of codewords $\boldsymbol{x}_i$ for which $-\ln P(\boldsymbol{y}|\boldsymbol{x}_i) = d$, and $\mathcal{D}_n$ is the set of values that the function $d(\boldsymbol{x}, \boldsymbol{y}) = -\ln P(\boldsymbol{y}|\boldsymbol{x})$ can take on for a given $\boldsymbol{y}$, as $\boldsymbol{x}$ exhausts the codebook $\mathcal{C}$. Note that as $d(\boldsymbol{x}, \boldsymbol{y})$ depends only on the empirical joint distribution of $\boldsymbol{x}$ and $\boldsymbol{y}$, then $|\mathcal{D}_n|$ cannot exceed the number of empirical conditional distributions (or conditional type classes) corresponding to pairs of n–sequences, and so, $|\mathcal{D}_n|$ is upper bounded by a polynomial in n.

Now, when a random code is considered, then instead of applying Jensen's inequality for $\beta \geq 1$ (as was done in [33]), and thereby insert the expectation operator into the square brackets, let us adopt another approach. Consider the following events:

$$\mathcal{B} = \left\{ \mathcal{C} : \; N_{\boldsymbol{y}}(d) \geq \exp\{n[R - \ln|\mathcal{X}| + h_0(d/n|\boldsymbol{y})]_+ + \epsilon]\} \text{ for some } d \in \mathcal{D}_n \right\},$$

where $[t]_+ \stackrel{\Delta}{=} \max\{0, t\}$ and where $h_0(\delta|\boldsymbol{y})$ is defined as the maximum of $H_Q(X|Y)$ subject to the constraints that $\boldsymbol{E}_Q\{d(X,Y)\} = \delta$ and that Y is distributed according to $\hat{P}_{\boldsymbol{y}}$. Also, define

$$\mathcal{W}_i = \left\{ \mathcal{C} : \; \min\{d : \; N_{\boldsymbol{y}}(d) \geq 1\} = i \right\}, \quad i \leq d_0(\boldsymbol{y}) \stackrel{\Delta}{=} n\delta_Y(R),$$

where we recall that $\delta_Y(R)$ is the solution to the equation $h_0(\delta|\boldsymbol{y}) = \ln|\mathcal{X}| - R$. Note that $\{\mathcal{W}_i\}$ are disjoint events. Now, for $\beta > \beta_c(R)$:

$$\begin{aligned}
& \boldsymbol{E} \left\{ \left[\sum_{d \in \mathcal{D}_n} N_{\boldsymbol{y}}(d) \cdot e^{-\beta d} \right]^{1/\beta} \right\} \\
\leq \; & \Pr\{\mathcal{B}\} \cdot [e^{nR} \cdot e^{-\beta \cdot 0}]^{1/\beta} + \\
& + \sum_{d \leq d_0(\boldsymbol{y})} \Pr\{\mathcal{W}_d \cap \mathcal{B}^c\} \cdot \left[e^{n\epsilon} e^{-\beta d} \right]^{1/\beta} + \\
& + \Pr\{\mathcal{W}_0^c \cap \mathcal{W}_1^c \cap \ldots \cap \mathcal{W}_{d_0(\boldsymbol{y})}^c \cap \mathcal{B}^c\} \cdot e^{-nF_g(Y)} \cdot e^{n\epsilon/\beta},
\end{aligned} \tag{17}$$

This inequality calls for some explanation: We are dividing the set of configurations of the RV's $\{N_{\boldsymbol{y}}(d)\}_{d \in \mathcal{D}_n}$ into three classes, defined by the events $\mathcal{B}$ and $\{\mathcal{W}_i\}$. In the first class, corresponding to the first term on the right–hand side, $\{N_{\boldsymbol{y}}(d)\}$ fall in $\mathcal{B}$, where there is at least one value

of d for which $N_{\boldsymbol{y}}(d)$ is *exponentially* larger (by at least ϵ) than its expectation. We bound the value of $[\sum_{d\in\mathcal{D}_n} N_{\boldsymbol{y}}(d)\cdot e^{-\beta d}]^{1/\beta}$, in this class, very "generously", by the maximum possible value it can possibly take, that is, when all e^{nR} codewords are at zero distance from $\boldsymbol{y}$, but this quantity is weighted by $\Pr\{\mathcal{B}\}$, which as is shown in the Appendix (Subsection A.2), decays double–exponentially rapidly, at least as fast as $e^{-e^{n\epsilon}}$, and so this first term is negligible. The other two classes correspond to $\mathcal{B}^c$, where for all $d\in\mathcal{D}_n$, $N_{\boldsymbol{y}}(d)$ does not exceed its expectation times $e^{n\epsilon}$. Here we distinguish between two cases (corresponding to the two other classes): In one of them, (at least) one of the distances below the generalized GV distance $d_0(\boldsymbol{y}) = n\delta_Y(R)$ is populated by subexponentially[8] many codewords. Since we are operating in the glassy regime, the dominant contribution to $[\sum_{d\in\mathcal{D}_n} N_{\boldsymbol{y}}(d)\cdot e^{-\beta d}]^{1/\beta}$ will be due to these minimum distance codewords, and the weighting of the event of minimum distance d is, of course, according to $\Pr\{\mathcal{W}_d\cap\mathcal{B}^c\}$. In the other case, which captures most of the probability mass (since it is the typical configuration of $\{N_{\boldsymbol{y}}(d)\}$), none of the distances below the generalized GV distance is populated by codewords, whereas for larger distances, $\{N_{\boldsymbol{y}}(d)\}$ are all (within a factor of $e^{n\epsilon}$) about their expectations. In this case, our expression again behaves according to the glassy regime, where the generalized GV distance dominates the partition function.

Now, regarding the second term, for $\delta = d/n < \delta_Y(R)$,

$$\Pr\{\mathcal{W}_d\cap\mathcal{B}^c\} \le \Pr\{\mathcal{W}_d\} \le \Pr\{N_{\boldsymbol{y}}(d)\ge 1\}, \tag{18}$$

where the latter expression is shown (Appendix, Subsection A.2) to decay at the exponential rate

[8]The event $\mathcal{B}^c$ guarantees that there are only subexponentially many codewords at distances below $d_0(\boldsymbol{y})$.

of $e^{-n[\ln|\mathcal{X}|-R-h_0(\delta|\boldsymbol{y})]}$. Thus,

$$\begin{aligned}
& \boldsymbol{E}\left\{\left[\sum_{d\in\mathcal{D}_n} N_{\boldsymbol{y}}(d)\cdot e^{-\beta d}\right]^{1/\beta}\right\} \\
\le\ & e^{-e^{n\epsilon}}\cdot[e^{nR}]^{1/\beta}+ \\
& +\sum_{\delta\le\delta_Y(R)} e^{-n[\ln|\mathcal{X}|-R-h_0(\delta|\boldsymbol{y})]}\cdot\left[e^{n\epsilon}e^{-\beta n\delta}\right]^{1/\beta}+e^{-nF_g(Y)}\cdot e^{n\epsilon/\beta} \\
\doteq\ & e^{n(R-\ln|\mathcal{X}|)}\cdot e^{n\epsilon/\beta}\exp\{n\max_{\delta\le\delta_Y(R)}[h_0(\delta|\boldsymbol{y})-\delta]\}+e^{-nF_g(Y)}\cdot e^{n\epsilon/\beta} \\
\doteq\ & e^{n(R-\ln|\mathcal{X}|)}\cdot e^{n\epsilon/\beta}\exp\{n[h_0(\delta_Y(R)|\boldsymbol{y})-\delta_Y(R)]\}+e^{-nF_g(Y)}\cdot e^{n\epsilon/\beta} \\
\doteq\ & e^{n(R-\ln|\mathcal{X}|)}\exp\{n[\ln|\mathcal{X}|-R-\delta_Y(R)]\}\cdot e^{n\epsilon/\beta}+e^{-nF_g(Y)}\cdot e^{n\epsilon/\beta} \\
\doteq\ & e^{-nF_g(Y)}\cdot e^{n\epsilon/\beta}.
\end{aligned} \tag{19}$$

Since ϵ can be chosen arbitrarily small for large n (in fact, one may let ϵ vanish with n sufficiently slowly), the exponential rate of the expression under discussion is actually bounded by $e^{-nF_g(Y)}$. Note that whenever $\beta\ge\beta_c$, this expression no longer depends on β. Finally, substituting this bound back into the bound on $\bar{P}_c$, we get:

$$\begin{aligned}
\bar{P}_c &\le \frac{1}{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n} e^{-nF_g(Y)} \\
&\doteq e^{-nR}\cdot e^{n\max_Y[H(Y)-F_g(Y)]} \\
&= e^{-n(R-\max_Y[H(Y)-F_g(Y)])}.
\end{aligned} \tag{20}$$

This calculation can be shown to be exponentially tight: a lower bound can be obtained by confining the calculation to the (high probability) event $\mathcal{W}_0^c\cap\mathcal{W}_1^c\cap\ldots\cap\mathcal{W}_{d_0(\boldsymbol{y})}^c\cap\mathcal{B}^c$ with the additional restriction that $N_{\boldsymbol{y}}(d)\ge\boldsymbol{E}\{N_{\boldsymbol{y}}(d)\}\cdot e^{-n\epsilon}$ for all $d\ge d_0(\boldsymbol{y})$ (i.e., the last term only in the above derivation). Note that in Arimoto's paper, where Jensen's inequality is used, the expectation of $\sum_d N_{\boldsymbol{y}}(d)e^{-\beta d}$ is computed, and this actually corresponds to the paramagnetic regime (without the constraint $H_Q(X|Y)\ge\ln|\mathcal{X}|-R$). The resulting bound might not be exponentially tight in general.[9] Finally, the optimization $\max_Y[H(Y)-F_g(Y)]$ can be carried out explicitly, yielding $\ln\sum_y e^{-f_g(y)}$, where where $f_g(y)=\boldsymbol{E}_{Q_{\beta_R}}\{d(X,Y)|Y=y\}$.

[9]Note that although the exact reliability function (for optimum codes) for rates above capacity was established by Dueck and Körner [34], here we are only focusing on random codes drawn under an i.i.d. distribution.

We have obtained then a random coding exponent formula in terms of the free energy density in the glassy phase, from which we learn that the free energy density of the glassy phase plays a central role in the calculation the exponent of correct decoding. To obtain some insight, it is instructive to examine this expression in the special case of the BSC. Here, since $F_g = F_g(Y)$ does not depend on the probability distribution of Y, we get:

$$\begin{aligned}\bar{P}_c &\leq e^{n[\ln 2 - R - F_g]} \\ &= e^{n[h(\delta_{GV}(R)) - \delta_{GV}(R)\ln\frac{1}{p} - (1-\delta_{GV}(R))\ln\frac{1}{1-p}]} \\ &= e^{-nD(\delta_{GV}(R)\|p)},\end{aligned} \tag{21}$$

where for $a, b \in (0,1)$, $D(a\|b) \stackrel{\Delta}{=} a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b}$. This result has the intuitively appealing interpretation of the probability of the large deviations event that the channel makes $n\delta_{GV}(R)$ errors or less, although $p > \delta_{GV}(R)$), in which case the correct codeword 'penetrates' into the sphere of radius $n\delta_{GV}(R)$, whose surface is populated by the codewords that dominate the glassy phase. Of course, when such an event happens, the correct codeword dominates the partition function, and thus the decoding is correct.

4 The Random Coding Error Exponent

Let us now examine rates below $I(X;Y)$. Consider Gallager's upper bound on the error probability for a given code [35]:

$$P_e \leq \frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n} P(\boldsymbol{y}|\boldsymbol{x}_m)^{1/(1+\rho)}\cdot\left[\sum_{m'\neq m} P(\boldsymbol{y}|\boldsymbol{x}_{m'})^{1/(1+\rho)}\right]^{\rho} \qquad \rho \geq 0. \tag{22}$$

The bracketed term is once again identified with $Z_e(\beta)$ for $\beta = \frac{1}{1+\rho} \leq 1$, in contrast to the calculation of P_c, where we used large values of β. For each m, let us first take only the expectation w.r.t. the incorrect codewords, referring to the random variables $\{N_{\boldsymbol{y}}(d)\}$. Let this partial expectation be denoted by $\tilde{P}_e$. We will also denote $\frac{1}{1+\rho}$ by β. One way to carry out this calculation is to use the same technique as we used in the previous section, by classifying the distance spectrum $\{N_{\boldsymbol{y}}(d)\}$ to its various classes. However, here since we know already that the use of Jensen's inequality would not harm the exponential tightness [36], it will be simpler to apply Jensen's inequality (for $0 \leq \rho \leq 1$, that is, $0.5 \leq \beta \leq 1$) and thereby essentially carry out the calculation in

the *paramagnetic* regime. We proceed then as follows:

$$\begin{aligned}
\tilde{P}_e &\le \frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n} P(\boldsymbol{y}|\boldsymbol{x}_m)^\beta \cdot \boldsymbol{E}\left\{\left[\sum_{d\in\mathcal{D}_n} N_{\boldsymbol{y}}(d)e^{-\beta d}\right]^\rho\right\} \\
&\le \frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n} P(\boldsymbol{y}|\boldsymbol{x}_m)^\beta \cdot \left[\sum_{d\in\mathcal{D}_n} \boldsymbol{E}\{N_{\boldsymbol{y}}(d)\}\cdot e^{-\beta d}\right]^\rho \\
&\doteq \frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n} P(\boldsymbol{y}|\boldsymbol{x}_m)^\beta \cdot \left[\sum_{d\in\mathcal{D}_n} e^{n[R-\ln|\mathcal{X}|+h_0(\delta|\boldsymbol{y})]}\cdot e^{-\beta d}\right]^\rho \\
&\doteq \frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n} P(\boldsymbol{y}|\boldsymbol{x}_m)^\beta \cdot [e^{-n\beta F_p(\beta,Y)}]^\rho. \qquad (23)
\end{aligned}$$

Next, we take the expectation w.r.t. the correct codeword $\boldsymbol{x}_m$: Define

$$\Gamma(y) = \ln\sum_{x\in\mathcal{X}} P^\beta(y|x) - \ln|\mathcal{X}|, \quad y\in\mathcal{Y}.$$

Then, the average error probability $\bar{P}_e$ is upper bounded by

$$\begin{aligned}
\bar{P}_e &\le \sum_{\boldsymbol{y}\in\mathcal{Y}^n} e^{\sum_{i=1}^n \Gamma(y_i)}\cdot e^{-n\rho\beta F_p(\beta,Y)} \\
&= \sum_{\boldsymbol{y}\in\mathcal{Y}^n} \exp\{n[\hat{\boldsymbol{E}}_{\boldsymbol{y}}\Gamma(y) - \rho\beta F_p(\beta,Y)]\} \\
&\doteq \exp\{n\cdot\max_Y[H(Y)+\sum_{y\in\mathcal{Y}}P(y)\Gamma(y)-\rho\beta F_p(\beta,Y)]\} \\
&= \exp\{-n\cdot\min_Y[\rho\beta F_p(\beta,Y)-\sum_{y\in\mathcal{Y}}P(y)\Gamma(y)-H(Y)]\}. \qquad (24)
\end{aligned}$$

Note that $\Gamma(y)$ is also related to a free energy expression, corresponding to the uniform prior over the entire input space $\mathcal{X}^n$, not only the codebook. Thus, we have two free energy expressions, one pertaining to the contribution of the correct codeword, and the other is related to the contributions of the incorrect codewords.

In the special case of the BSC, where $F_p(\beta,Y)$ and does not depend on Y and $\Gamma=\Gamma(y)$ does

not depend on y, we get the exponential rate of

$$\begin{aligned}
& \min_{Y}[\beta\rho F_p(\beta) - \Gamma - H(Y)] \\
= & \ \beta\rho F_p(\beta) - (\ln[p^\beta + (1-p)^\beta] - \ln 2) - \ln 2 \\
= & \ \rho(\ln 2 - R) - (1+\rho)\ln[p^\beta + (1-p)^\beta] \\
= & \ \rho \ln 2 - (1+\rho)\ln[p^{1/(1+\rho)} + (1-p)^{1/(1+\rho)}] - \rho R \\
= & \ E_0(\rho) - \rho R \\
\stackrel{\Delta}{=} & \ E_0(\rho, R)
\end{aligned} \tag{25}$$

which is, as expected, Gallager's reliability function for the BSC. The optimum choice of ρ depends on R. As is shown in [35, pp. 151-152], in the range $R \leq \ln 2 - h(p_{1/2})$, that is, $p_{1/2} < \delta_{GV}(R)$, we have $\rho = 1$, which means $\beta = \frac{1}{2}$. For $R \in [\ln 2 - h(p_{1/2}), \ln 2 - h(p)]$, the optimum ρ is in $[0, 1)$, and it satisfies $R = \ln 2 - h(p_{1/(1+\rho)}) = \ln 2 - h(p_\beta)$, or, equivalently, $p_\beta = \delta_{GV}(R)$, which means that we move along the boundary between the the glassy phase and the paramagnetic phases of $Z_e(\beta|\boldsymbol{y})$.

Appendix

A.1 Proof of the Concavity of $J_Y(\beta,\cdot)$

Let Q_1 and Q_2 achieve $J_Y(\beta,R_1)$ and $J_Y(\beta,R_2)$, respectively. Now, let $Q=\alpha Q_1+(1-\alpha)Q_2$ for some $\alpha\in(0,1)$. First, observe that by the concavity of the conditional entropy in $Q_{X|Y}$ for fixed Q_Y, we have

$$H_Q(X|Y)\geq \alpha H_{Q_1}(X|Y)+(1-\alpha)H_{Q_2}(X|Y)\geq \ln|\mathcal{X}|-\alpha R_1-(1-\alpha)R_2.$$

It follows then that $H_Q(X|Y)-\beta\boldsymbol{E}_Q d(X,Y)\leq J(\beta,\alpha R_1+(1-\alpha)R_2|\boldsymbol{y})$. But, on the other hand

$$\begin{aligned} H_Q(X|Y)-\beta\boldsymbol{E}_Q d(X,Y) &\geq \alpha[H_{Q_1}(X|Y)-\beta\boldsymbol{E}_{Q_1}d(X,Y)]+(1-\alpha)[H_{Q_2}(X|Y)-\beta\boldsymbol{E}_{Q_2}d(X,Y)] \\ &= \alpha J_Y(\beta,R_1)+(1-\alpha)J_Y(\beta,R_2). \end{aligned} \tag{26}$$

Thus,

$$J_Y(\beta,\alpha R_1+(1-\alpha)R_2)\geq \alpha J_Y(\beta,R_1)+(1-\alpha)J_Y(\beta,R_2).$$

A.2 Large Deviations Behavior of $N_{\boldsymbol{y}}(d)$

For $a,b\in[0,1]$, consider the binary divergence

$$\begin{aligned} D(a\|b) &\stackrel{\Delta}{=} a\ln\frac{a}{b}+(1-a)\ln\frac{1-a}{1-b} \\ &= a\ln\frac{a}{b}+(1-a)\ln\left[1+\frac{b-a}{1-b}\right] \end{aligned} \tag{27}$$

To derive a lower bound to $D(a\|b)$, let us use the inequality

$$\ln(1+x)=-\ln\frac{1}{1+x}=-\ln\left(1-\frac{x}{1+x}\right)\geq\frac{x}{1+x},$$

and then

$$\begin{aligned} D(a\|b) &\geq a\ln\frac{a}{b}+(1-a)\cdot\frac{(b-a)/(1-b)}{1+(b-a)/(1-b)} \\ &= a\ln\frac{a}{b}+b-a \\ &> a\left(\ln\frac{a}{b}-1\right). \end{aligned} \tag{28}$$

Now, let $N_{\boldsymbol{y}}(d)$ denote the number of codewords for which $-\ln P(\boldsymbol{y}|\boldsymbol{x}_i)=d$. As mentioned earlier, $N_{\boldsymbol{y}}(d)$ is the sum of the e^{nR} independent binary random variables $1\{d(\boldsymbol{X}_i,\boldsymbol{y})=d\}$, where the

probability that $d(\boldsymbol{X}_i, \boldsymbol{y}) = d$ is exponentially $b = e^{-n[\ln|\mathcal{X}| - h_0(\delta|\boldsymbol{y})]}$, $h_0(\delta|\boldsymbol{y})$ being the maximum of $H_Q(X|Y)$ subject to the constraints that $\boldsymbol{E}_Q\{d(X,Y)\} = \delta$, $\delta = d/n$, and that Y is distributed according to $\hat{P}_{\boldsymbol{y}}$. The event $N_{\boldsymbol{y}}(d) \geq e^{nA}$, for $d = \delta n$ and $A \in [0, R)$, means that the relative frequency of the event $1\{d(\boldsymbol{X}_i, \boldsymbol{y}) = d\}$ is at least $a = e^{-n(R-A)}$. Thus, by the Chernoff bound:

$$\begin{aligned}
\Pr\{N_{\boldsymbol{y}}(d) \geq e^{nA}\} &\leq \exp\left\{-e^{nR} D(e^{-n(R-A)} \| e^{-n[\ln|\mathcal{X}| - h_0(\delta|\boldsymbol{y})]})\right\} \\
&\leq \exp\left\{-e^{nR} \cdot e^{-n(R-A)} (n[(\ln|\mathcal{X}| - R - h_0(\delta|\boldsymbol{y}) + A] - 1)\right\} \\
&\leq \exp\left\{-e^{nA} (n[\ln|\mathcal{X}| - R - h_0(\delta|\boldsymbol{y}) + A] - 1)\right\}. \qquad (29)
\end{aligned}$$

Now, for $A = [R - \ln|\mathcal{X}| + h_0(\delta|\boldsymbol{y})]_+ + \epsilon$, the term in the square brackets is at least $\epsilon > 0$, and thus $\Pr\{N_{\boldsymbol{y}}(d) \geq e^{nA}\}$ decays double–exponentially rapidly, not slower than $e^{-e^{n\epsilon}}$. The probability of the union of the (polynomially many) events $\{N_{\boldsymbol{y}}(d) \geq e^{nA}\}_{d \in \mathcal{D}_n}$, which is upper bounded by the sum of the probabilities, is still double–exponentially small. Thus, $\Pr\{\mathcal{B}\}$ decays double–exponentially rapidly. Now, the event $\{N_{\boldsymbol{y}}(d) \geq 1\}$ corresponds to the choice $A = 0$. For $\delta < \delta_Y(R)$, $\delta_Y(R)$ being the solution to the equation $\ln|\mathcal{X}| - R = h_0(\delta|\boldsymbol{y})$, which means that $\ln|\mathcal{X}| - R - h_0(\delta|\boldsymbol{y}) > 0$, this gives an ordinary exponential decay at the rate of $e^{-n[\ln|\mathcal{X}| - R - h_0(\delta|\boldsymbol{y})]}$.

References

[1] I. Kanter and D. Saad, "Error–correcting codes that nearly saturate Shannon's bound," *Physical Review Letters*, vol. 83, no. 13, pp. 2660–2663, September 1999.

[2] Y. Kabashima, N. Sazuka, K. Nakamura, and D. Saad, "Tighter decoding reliability bound for Gallager's error–correcting code," *Physical Review E*, vol. 64, pp. 046113-1–046113-4, 2001.

[3] N. Sourlas, "Spin–glass models as error–correcting codes," *Nature*, pp. 693–695, vol. 339, June 1989.

[4] N. Sourlas, "Spin glasses, error–correcting codes and finite–temperature decoding," *Europhysics Letters*, vol. 25, pp. 159–164, 1994.

[5] Y. Kabashima and D. Saad, "Statistical mechanics of error correcting codes," *Europhysics Letters*, vol. 45, no. 1, pp. 97–103, 1999.

[6] D. Guo and S. Verdú, "Multiuser detection and statistical physics," preprint, August 2002.

[7] O. Shental, I. Kanter, and A. J. Weiss, "Capacity of complexity-constrained noise-free CDMA," preprint 2005.

[8] T. Tanaka, "A statistical–mechanics approach to large–system analysis of CDMA multiuser detectors," *IEEE Trans. Inform. Theory*, vol. 48, no. 11, pp. 2888–2910, November 2002.

[9] T. Tanaka, "Statistical mechanics of CDMA multiuser demodulation," *Europhysics Letters*, vol. 54, no. 4, pp. 540–546, 2001.

[10] A. Procacci and B. Scoppola, "Statistical mechanics approach to coding theory," *J. of Statistical Physics*, vol. 96, nos. 3/4, pp. 907–912, 1999.

[11] T. Hosaka and Y. Kabashima, "Statistical mechanical approach to error exponents of lossy data compression," *J. Physical Society of Japan*, vol. 74, no. 1, pp. 488–497, January 2005.

[12] Y. Kabashima, K. Nakamura, and J. van Mourik, "Statistical mechanics of typical set decoding," *Physical Review E*, vol. 66, 2002.

[13] N. Sarshar and X. Wu, "Statistical mechanics of optimal networked source coding," preprint.

[14] T. Mutayama, "Statistical mechanics of the data compression theorem," *J. Phys. A: Math. Gen.*, vol. 35, pp. L95–L100, 2002.

[15] I. Rojdestvenski and M. C. Cottman, "Mapping of statistical physics to information theory-with appication to biological systems," *J. Theor. Biol.*, pp. 43–54, 2000.

[16] O. Shental and I. Kanter, "Shannon capacity of infinite–range spin–glasses," preprint 2006.

[17] T. Mora and O. Rivoire, "Statistical mechanics of error exponents for error–correcting codes," arXiv:cond-mat/0606696, June 2006.

[18] Y. Kabashima and T. Hosaka, "Statistical mechanics for source coding with a fidelity criterion," *Progress of Theoretical Physics*, Supplement no. 157, pp. 197–204, 2005.

[19] M. Mézard and A. Montanari, *Constraint satisfaction networks in physics and computation*, draft, February 27, 2006. Available on–line at: [http://www.stanford.edu∼montanar/BOOK/book.html].

[20] D. McAllester, "A statistical mechanics approach to large deviations theorems," preprint, 200?. Available on-line at: [http://citeseer.ist.psu.edu/443261.html].

[21] Y. Oono, "Large deviation and statistical physics," *Progress of Theoretical Physics Supplement*, no. 99, pp. 165–205, 1989.

[22] P. T. Landsberg and V. Vedral, "Distributions and channel capacities in generalized statistical mechanics," *Physics Letters A*, 247, pp. 211–217, October 1998.

[23] A. Montanari, "Two lectures on iterative decoding and statistical mechanics," arXiv:cond-mat/0512296, December 14, 2005.

[24] A. Montanari and R. Urbanke, "Modern coding theory: the statistical mechanics and conputer science point of view," preprint 2007.

[25] A. Montanari, "The glassy phase of Gallager codes," arXiv:cond-mat/0104079v1, April 4, 2001.

[26] P. Ruján, "Finite temperature error–correcting codes," *Phys. Rev. Let.*, vol. 70, no. 19, pp. 2968–2971, May 1993.

[27] S. Franz, M. Leone, A. Montanari, and F. Ricci–Tersenghi, "The dynamic phase transition for decoding algorithms," arXiv:cond-mat/0205051v1, May 2, 2002 (also in Phys. Rev. E 66, 2002).

[28] K. Tadaki, "A statistical mechanical interpretation of instantaneous codes," *Proc. ISIT 2007*, pp. 1906–1910, Nice, France, June 2007.

[29] Y. Iba, "The Nishimori line and Bayesian statistics," *J. Phys. A: Math. Gen.*, vol. 32, pp. 3875–3888, 1999.

[30] B. Derrida, "Random–energy model: an exactly solvable model for disordered systems," *Phys. Rev. B*, vol. 24, no. 5, pp. 2613–2626, September 1981.

[31] J. Ziv, "Universal decoding for finite–state channels," *IEEE Trans. Inform. Theory*, vol. IT-31, no. 4, pp. 435–460, July 1985.

[32] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, 1981.

[33] S. Arimoto, "On the converse to the coding theorem for discrete memoryless channels," *IEEE Trans. Inform. Theory*, vol. IT-19, no. 5, pp. 357–359, May 1973.

[34] G. Dueck and J. Körner, "Reliability function of a discrete memoryless channel at rates above capacity," *IEEE Trans. Inform. Theory*, vol. IT-25, no. 1, pp. 82–85, January 1979.

[35] A. J. Viterbi and J. K. Omura, *Principles of digital communication and coding*, McGraw Hill, New York, 1979.

[36] R. G. Gallager, "The random coding bound is tight for the average code," *IEEE Trans. Inform. Theory*, vol. IT–19, no. 2, pp. 244–246, March 1973.